Probabilidad y Estadística

13

Actividades de Aprendizaje

Conceptos y definiciones de esta clase:

- Interpretación del error estándar
- Variación Explicada, No Explicada y Total
- El Coeficiente de Determinación
- El Coeficiente de Correlación
- Introducción al Análisis de Regresión Múltiple

1.1 Interpretación del error estándar

Vamos a interpretar el error calculado en el punto anterior. En primer lugar, indicaremos que un valor de S_e cercano a 0 indica que los puntos del diagrama de dispersión se ajustan adecuadamente a la recta obtenida. Por el contrario, un valor cercano a 1 nos informa que los puntos se hallan muy dispersos respecto de la recta.

Pero, además, si consideramos que los puntos tienen una distribución normal respecto de la recta, el análisis de la distribución nos permite encontrar los valores extremos de la recta para los cuales un amplio porcentaje de los mismos quedan encerrados entre dichos valores. Por ejemplo:

- Un valor de $\pm 1S_e$ encierra el 68.2% de los puntos
- Un valor de $\pm 2S_e$ encierra el 95.4% de los puntos
- Un valor de $\pm 3S_e$ encierra el 99.6% de los puntos

Lo anterior, aplicado a la recta que hemos encontrado, se manifiesta de la siguiente manera:

- Para $\pm 1S_e$ las rectas son $y = 3.144\,x - 0.05 \pm 0.093$
- Un valor de $\pm 2S_e$ las rectas son $y = 3.144\,x - 0.05 \pm (2)(0.093)$
- Un valor de $\pm 3S_e$ las rectas son $y = 3.144\,x - 0.05 \pm (3)(0.093)$

Como en nuestro caso el error es muy bajo, una pequeña variación en la altura de la recta incluye un alto porcentaje de los puntos del diagrama.

1.2 Variación Explicada, No Explicada y Total

En esta sección mencionaremos algunos parámetros que forman parte del análisis de la variabilidad en el análisis de regresión.

A la diferencia entre cada valor $\hat{y}_i$ estimado mediante nuestra recta de regresión y la media $\bar{y}$ se lo denomina ***variación explicada*** de la variable de respuesta. Se la denomina de esta manera porque son variaciones que fueron consideradas por la recta de regresión. En símbolos:

$$\text{variación explicada de } y_i = \hat{y}_i - \bar{y}$$

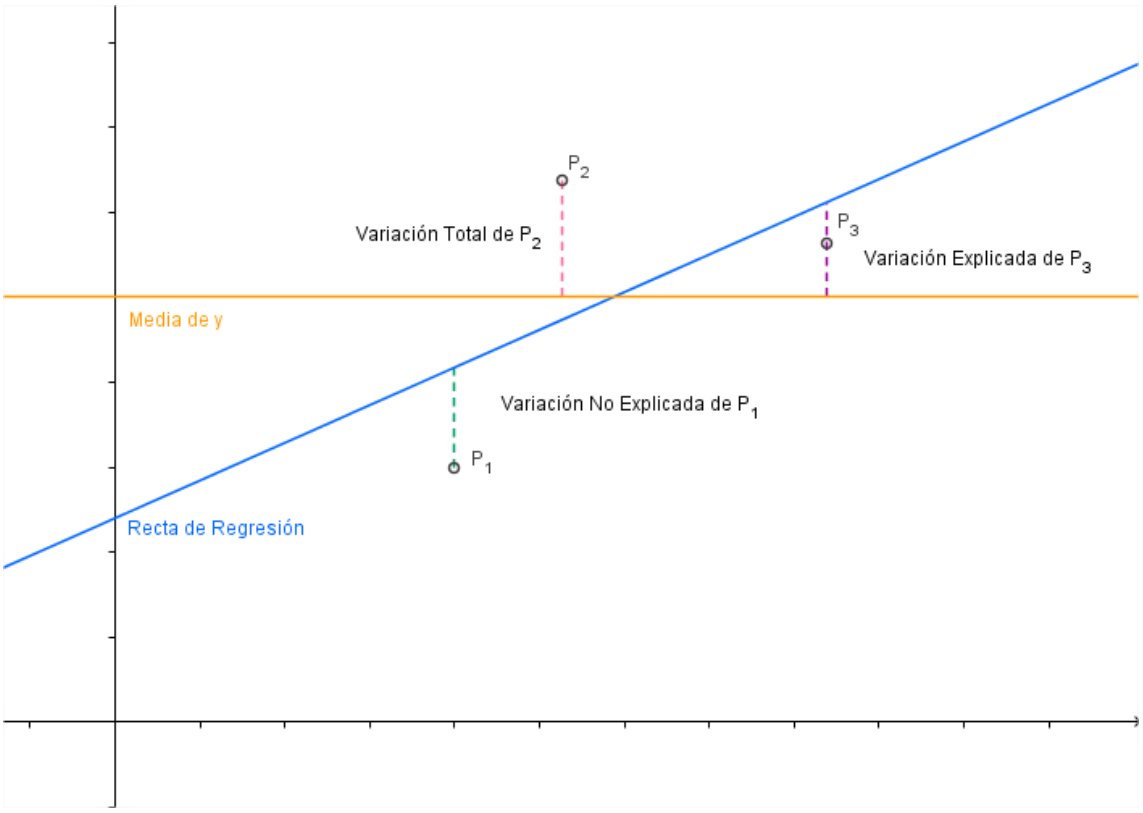
También encontramos diferencias entre los valores observados y_i y los valores calculados a través de la recta de regresión. Estas son las denominada variaciones no explicadas, que se deben a factores no considerados por la recta de regresión. En símbolos:

$$\text{variación no explicada de } y_i = y_i - \hat{y}_i$$

Finalmente, a la diferencia entre cada valor de y_i obtenido para nuestro análisis y la media se lo denomina variación total. En símbolos:

$$\text{variación total de } y_i = y_i - \bar{y}$$

En la siguiente gráfica se muestran los tres tipos de variaciones aplicados a tres puntos distintos.



En la gráfica se aprecia claramente entre qué objetos se establece la diferencia que figura en cada variación, de la siguiente manera:

La variación…	es la diferencia entre…
no explicada	*el valor observado y la regresión*
explicada	*el valor de regresión y la media*
total	*el valor observado y la media*

1.3 El Coeficiente de Determinación

Una duda que se desprende del análisis de regresión realizado es en qué medida la variable independiente "explica" a la variable dependiente mediante la recta de regresión.

Esto puede medirse mediante el coeficiente de determinación r^2 que muestra el porcentaje de la variación de y que explica nuestro modelo.

Se puede calcular con la siguiente fórmula:

$$r^2 = 1 - \frac{variación\ de\ los\ valores\ de\ y\ respecto\ de\ la\ recta\ de\ regresión}{variación\ de\ los\ valores\ de\ y\ respecto\ de\ su\ media\ aritmética}$$

O sea,

$$r^2 = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2}$$

Calculemos este estimador para nuestro ejemplo y realicemos su análisis. Para ello, se muestra a continuación la tabla con los valores originales y los que necesitamos para realizar nuestros cálculos.

Nro. de observación (n)	Diámetro en cm (x)	Longitud en cm (y)	$\hat{y}$	$(y_i - \hat{y}_i)^2$	$(y_i - \bar{y})^2$
1	2.10	6.50	6.55	0.00262	44.89
2	5.50	17.10	17.24	0.02050	15.21
3	4.00	12.50	12.52	0.00068	0.49
4	3.80	12.00	11.89	0.01056	1.44
5	6.00	18.90	18.81	0.00713	32.49
6	3.50	11.00	10.95	0.00213	4.84
7	4.60	14.40	14.41	0.00016	1.44
Totales	29.50	92.40	92.40	0.04381	100.8

$$r^2 = 1 - \frac{\sum_1^n (y_i - \hat{y}_i)^2}{\sum_1^n (y_i - \bar{y})^2} = 1 - \frac{-7.105 \text{ x } 10^{-15}}{100.8} \cong 1$$

Como el valor de este resultado es prácticamente 1, entonces el modelo explica el 100% de las variaciones de las variables de respuesta. Esto ocurre porque todos los valores caen sobre la recta de regresión o son muy próximos a ella.

Una vez más, como en casos anteriores, aportamos una fórmula alternativa para encontrar el coeficiente de determinación, y que no implica realizar los cálculos de las nuevas columnas. La fórmula para el método abreviado es:

$$R^2 = \frac{b\sum_1^n x_i y_i + a\sum_1^n y_i - n\bar{y}^2}{\sum_1^n {y_i}^2 - n\bar{y}^2}$$

Que para los valores de nuestro ejemplo quedaría como:

$$r^2 = \frac{(3.144)(421.44) + (-0.05)(92.4) - (7)(13.2)^2}{1320.48 - (7)(13.2)^2} \cong 1$$

Puesto en términos de las variaciones estudiadas en el punto anterior, podemos decir que el coeficiente de determinación es la relación entre la variación explicada y la variación total, lo cual nos brinda una nueva fórmula para su cálculo, a saber:

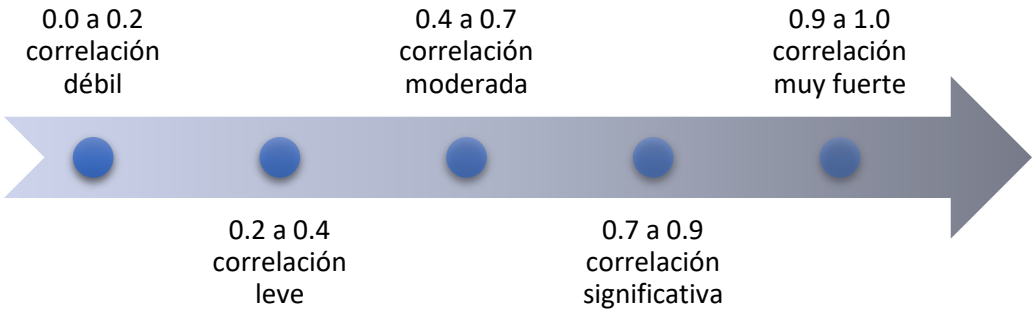
$$r^2 = \frac{\sum_1^n (\hat{y}_i - \bar{y})^2}{\sum_1^n (y_i - \bar{y})^2}$$

1.4 El Coeficiente de Correlación

El coeficiente de correlación nos indica qué tan fuerte (o débil) es la relación entre las variables analizadas, y su valor es la raíz cuadrada del coeficiente de determinación, si bien puede calcularse independientemente de él. Su valor está siempre comprendido entre -1 y 1, siendo su signo el que le corresponde a la pendiente de la recta de regresión. Tenemos entonces que:

$$r = \sqrt{r^2}$$

Y su interpretación



Interpretación conjunta de ambos coeficientes

Si el coeficiente de determinación es 0.64 entonces el coeficiente de correlación es de 0.80 y se interpreta de la siguiente manera:

"*El 64% de la variación en la variable dependiente queda explicada por la recta de regresión, y un 80% de los datos están relacionados entre sí*"

Como la relación entre ambos coeficientes siempre será la misma independientemente del tema tratado, las decisiones estratégicas basadas en estos indicadores deberán estar más orientadas a lo abarcativo o a lo efectivo, según demos mayor importancia al coeficiente de correlación o al de determinación, respectivamente.

1.5 Introducción al Análisis de Regresión Múltiple

Los conceptos vistos en los puntos anteriores pueden ampliarse al caso en el que los resultados obtenidos para la variable de respuesta dependan de más de una variable independiente. En estos casos, al análisis lo denominamos de Regresión Múltiple.

Los ejemplos de este tipo apuntan a describir el comportamiento de alguna variable cuando se sospecha que son influenciadas por más de motivo, como por ejemplo el rendimiento de un deportista por raza y procedencia, los costos de publicidad cuando se combinan métodos gráficos y audiovisuales, la combinación de diversas medicaciones o vitaminas en una dieta, los factores diversos que desencadenan una enfermedad, la estructura salarial de una compañía y otros por el estilo.

Para realizar el análisis de regresión múltiple se procede de manera similar a la que hemos utilizado para la regresión simple, esto es:

- Hallamos una ecuación que se ajuste a la regresión múltiple
- Encontramos el error de la regresión
- Analizamos los coeficientes que nos permitan conocer la correlación entre las variables

Comencemos entonces, por la ecuación para la regresión múltiple:

Así como la ecuación que define los valores estimados de la variable de respuesta es $y = a + bx$ cuando depende de una única variable independiente, cuando depende de muchas será:

$$y = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$$

donde x_1, x_2, x_3 ... son las variables predictoras y b_1, b_2, b_3 .... sus pendientes asociadas. Se trata de un modelo de regresión lineal múltiple con n regresores (o variables predictoras)

Y así como en el caso de única dependencia la solución gráfica es una recta de regresión, al agregar una variable predictora más, por ejemplo, se obtiene un plano de regresión. Pero, así y todo, el ajuste por mínimos cuadrados se realizará de la misma manera, es decir, minimizando las sumas de los cuadrados de los errores.

Nota:

Cuando una regresión múltiple responde a una fórmula del tipo $y = a + b_1x_1 + b_2x_2 + b_3x_3 + \cdots + b_nx_n$ con n parámetros ***lineales*** estaremos en presencia de un modelo de regresión lineal, independientemente del tipo de superficie que genere la función.